Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 00 (2012) 1–11
Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Visualizing Climate Variability with Time-Dependent Probability Density Functions, Detecting it with Information Theory

J. Walter Larson[a,b,c]

[a]*Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA*
[b]*Computation Institute, University of Chicago*
[c]*Research School of Computer Science, The Australian National University*

Abstract

A framework is presented for visualizing and detecting climate variability and change based on time-dependent probability density functions (PDFs). The PDFs show how the distribution of values in the sample window changes over time and show more detail than do timeseries of windowed moments. A set of information-theoretic statistics based on the Shannon entropy and the Kullback-Leibler divergence (KLD) are defined to assess PDF complexity and temporal variability. The KLD-based measures quantify the representativeness of a 30-year sampling window of a larger climatic record: how well a long sample can predict a smaller sample's PDF, and how well one 30-year sample matches a similar sample shifted in time. These information-theoretic statistics constitute a new type of climate variability, *informatic variability*. These techniques are applied to the Central England Temperature record, the longest continuous meteorological observational record.

Keywords: Probability Density Function, Information Theory, Climate Variability, Climate Change

1. Introduction

Climate is a statistical construct computed from meteorological state data sampled over a predefined period. By convention, this window sampling period W is 30 years—a number arrived at by a vote at the 1937 International Meteorological Organization meeting [1, 2]. Climate models do not model the climate directly; they compute solutions to equations of evolution for the Earth system's instantaneous state and write daily or monthly summaries of the state to history files, which are then postprocessed to compute climatologies. The I/O-intensive nature of coupled climate models is arguably the most significant barrier to creating an exascale climate model. A natural question arises: Can we model the climate directly? Hasselmann [3] proposed a statistical dynamical model (SDM) approach that used the Fokker-Planck equation (FPE) as the equation of evolution for the probability density function (PDF). For some quantity X and time t with time-dependent PDF $\rho(X,t)$, the univariate FPE is

$$\partial_t \rho(X,t) + \partial_X[D_1(X,t)\rho(X,t)] = \partial_X[D_2(X,t)\partial_X\rho(X,t)]. \quad (1)$$

In (1), $D_1(X,t)$ is the *drift* term, and $D_2(X,t)$ is the *diffusion* term. In Hasselmann's narrative, weather systems—atmospheric variations of duration on the order of 15 days or less—provided stochastic forcing to more slowly responding, integrative components of the Earth system (oceans, cryosphere, and biosphere), much in the way that molecular motions drive Brownian motion of larger particles suspended in a fluid. Direct deterministic modeling of

a time-dependent climate PDF using something like (1) is a potent idea because the PDF provides a comprehensive description of a sample's underlying population statistics. Direct dynamic modeling of $\rho(X,t)$ would, in theory, enable deterministic modeling of moments and estimates for quantiles and extrema. SDMs, however, fell out of favor a decade after Hasselman's 1976 paper because the role of atmospheric momentum transport was not adequately covered by SDMs and advances in computer hardware made practicable the general circulation model-based approach employed in current coupled climate and Earth system models.

An alternative to the SDM approach is to ask a slightly different question: What does the *empirical* climatic PDF $\rho(X,t)$ look like for the observational timeseries of some meteorological variable X, and is it possible to *fit* an equation of evolution similar to the FPE for $\rho(X,t)$? Here I tackle the first part of this question by posing—and proposing answers to—the following questions regarding a windowed sample of W years taken from a longer record of Y years:

Q1 What does the W-window-sampled $\rho(t,X)$ look like, how confident can we be of its structure, and how does it evolve in time?

Q2 What is the information content of a sample of W years? How does it evolve in time over the record of Y years, and what does this signify?

Q3 How well does a sample of W years predict the whole available data record of $Y > W$ years or other major climatically relevant subsets of the whole record?

Q4 How well does knowledge of the whole record's time-independent PDF $\rho(X)$ predict a local W-window-sample-generated PDF $\rho(t,X)$? That is, when viewed with prior knowledge of the parent density function, how unusual does $\rho(t,X)$ look?

Q5 How well does one W-window sample's PDF $q(t,X)$ predict another W-window sample's PDF $p(t',X)$?

Q6 Is it possible to use time-dependent PDFs to classify periods of time that are climatically stable or undergoing climatic change?

I will address **Q1** by using a density estimation technique that employs a Bayesian-derived optimal binning scheme. This binning scheme provides estimated PDFs in a form that is highly compatible with computing key information-theoretic statistics. These information-theoretic statistics provide the means to address **Q2–Q6** and constitute a new type of climatic variability—*informatic variability*. In particular, this approach expresses differences in PDFs from different climatic sampling periods—and, by association, climate change—as a form of information loss, specifically, loss of ability of a past (changed) climate record's PDF to predict a changed (past) climate record's PDF. Application of these techniques to a classic meteorological timeseries—the Central England Temperature (CET) record [4])—provide striking visualizations of the evolution of the climate over this record, reveals previously known properties, and puts in stark contrast the current climate's oddity with respect to the previous observational record.

2. Probability Density Functions, Information Theory, and Climate

For a random continuous variable $X \in (-\infty,\infty)$ the *probability density function* $p(X)$ satisfies the following conditions: $p(x) > 0, \forall x \in (-\infty,\infty)$, and $\int_{-\infty}^{\infty} p(x)dx = 1$. Suppose x depends on the time t. The dependency $x \to x(t)$ implies potential time dependence in the PDF; that is, $p(x) \to p(x,t)$. Note that each "time slice" of $p(x,t)$ satisfies the normalization condition of a univariate PDF; that is, for $t = t_C$, $\int_{-\infty}^{\infty} p(x,t)|_{t=t_C}dx = 1$. If the underlying statistics of X remain stationary, then the PDF remains solely a function of x. Nonstationarity—temporal sensitivity of the PDF—raises the question of how to estimate $p(x,t)$. A common technique used by the climate community is to sample $x(t)$ using a time window of width W and centered at a time t_C, resulting in a sample $\mathcal{S} = x(t), t \in [t_C - \frac{W}{2}, t_C + \frac{W}{2})$, and estimating a univariate $p(x)|_{t=t_C}$ to get $p(x,t_C)$. Advancing this window through time then provides $p(x,t)$. This windowed sampling technique underlies PDF estimation in this paper. The windowed sampling and binning technique described in Section 3, combined with visualization, answer **Q1**.

Information theory [5, 6] is a mathematical framework for quantifying information content and identifying relationships between random variables. It has been used extensively in the telecommunications and signal-processing

communities; but it also has been applied by the climate community to solve problems of predictability [7] and model-reality comparison [8, 9] and to evaluate climate sampling window sizes [10]. Its conceptual roots lie in Boltzmann's statistical mechanical formulation of thermodynamic entropy. The Shannon entropy (SE) $H(X)$ is

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx. \tag{2}$$

The logarithm base in (2)—and in (3) below—defines the units for information; for bases 2 and e, $H(X)$ is measured in *bits* and *nats*, respectively. $H(X)$ broadly quantifies the amount of "surprise" in the distribution of values of X. Given a precomputed PDF, computing the SE provides an answer to information content component of **Q2**. Note that the integral formulation of the SE (2) can yield negative or infinite values; the reason is that the probability density function $p(x)$ may locally exceed unity.

Consider two distinct PDFs for X: $p(x)$ and $q(x)$. The additional information, or *gain*, required to predict $p(x)$ given $q(x)$ is the *Kullback-Leibler divergence* (KLD):

$$D_{KL}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \left[\frac{p(x)}{q(x)} \right] dx. \tag{3}$$

For a time-dependent variable $x(t)$, similar arguments to those presented for PDFs can be applied to compute time-window-sampled SE and KLD values from (2) and (3), respectively.

The KLD quantifies differences between PDFs. Sometimes it is called a "distance measure" or "metric" for PDFs, but such terms are inaccurate: The KLD is not symmetric; in general, $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$. Furthermore, the KLD does not satisfy the triangle inequality. The nonsymmetric nature of the KLD is of particular use. The representativeness of a particular W-window's PDF of a larger record can be addressed by constructing $q(x,t)$ for each W-window of a climate record, using the entire Y-year record to construct a density $p(x)$, and computing $D_{KL}(p \parallel q)$. In fact, $D_{KL}(p \parallel q)$ provides an answer to **Q3**. Reversing the arguments, an answer to **Q4** is $D_{KL}(q \parallel p)$. **Q5** may be answered by constructing $p(x,t)$ from two different W-windows and computing the KLD. That is, one can construct two samples $\mathcal{S}_1 = x(t), t \in [t_1 - \frac{W}{2}, t_1 + \frac{W}{2})$ and $\mathcal{S}_2 = x(t), t \in [t_2 - \frac{W}{2}, t_2 + \frac{W}{2})$, and compute from them $p(x,t_1)$ and $p(x,t_2)$, respectively. The time-shifted KLD $D_{KL}(p(x,t_1) \parallel p(x,t_2))$ quantifies how much additional information is required to predict the PDF at $t = t_2$ from the PDF at $t = t_1$. Large-scale application of the time-shifted KLD to all possible W-sampling windows in a record of Y years can provide clues about periods of relative climatic stability and rapid climate change, providing answers to **Q6**.

The aforementioned SE and KLD statistics characterize the system's informatic variability.

Both the SE and KLD were originally formulated for discrete variables; one replaces the PDFs $\{p(x), q(x)\}$ with the *probability mass functions* (PMF) $\{\vec{\pi}, \vec{\zeta} \in \Re^N\}$, respectively, and integrals w.r.t. x in (2) and (3) with summations over N discrete states [6]. All PDFs plotted and computed SE and KLD quantities presented in this paper derived from PMFs estimated from sample data. Numerical computation of the SE is sensitive to the discretization $dx \to \Delta x$. Thus, one must take care in discretizing or binning continuous data to form a PMF or PDF, respectively.

3. Computational Methodology

PDF estimation from observational data is nontrivial and remains an active area of research; an excellent overview is given in [11]. I used an optimal binning scheme derived from Bayesian principles [12]. The technique's underlying assumptions are threefold: a discrete uniform prior distribution for the number of bins within a feasible range, a Dirichlet prior distribution for the bin probabilities, and a piecewise constant PDF with uniform bins. Other prior knowledge I comprises the data sample's size N, range V, and implied bin boundary locations assuming uniform widths. The posterior probability $P(M|\vec{d}, I)$ that a PDF with M bins describes the data sample $\vec{d}$ is [12]

$$P(M|\vec{d}, I) \propto \left(\frac{M}{V} \right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^{M} \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function and n_k is the number of counts in each bin. The value of M that maximizes $P(M|\vec{d}, I)$ yields the most probable piecewise constant, uniform-bin-width PDF for $\vec{d}$. Maximizing the logarithm of the right-hand side (RHS) of (4) is computationally easier and also maximizes $P(M|\vec{d}, I)$. The optimally estimated PDF is

parameterized as

$$\mu_i = \left(\frac{M}{V}\right)\left(n_i + \frac{1}{2}\right)\left(N + \frac{M}{2}\right)^{-1} \tag{5}$$

$$\sigma_i^2 = \left(\frac{M}{V}\right)^2\left(n_i + \frac{1}{2}\right)\left(N - n_i + \frac{M-1}{2}\right)\left(N + \frac{M}{2} + 1\right)^{-1}\left(N + \frac{M}{2}\right)^{-2}, \tag{6}$$

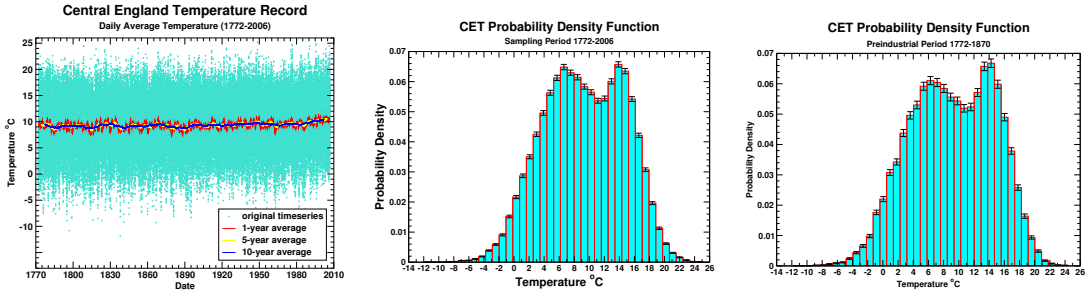
where μ_i and σ_i^2 are the bin probability densities and associated variances, respectively [12]. Even when data is absent from a bin, $\mu_i \neq 0$, and the μ_i are normalized by construction. The density and its associated uncertainties are defined by (5) and (6), respectively, thus answering **Q1**. Densities computed in this manner are used to evaluate all SE- and KLD-based statistics presented in this paper.

Another desirable characteristic of this PDF estimation scheme is that it can detect severely truncated data [13]. Truncated data will clump at its truncated values, adding fine structure to a PDF's broad shape, spawning a series of local maxima in the RHS of (4), growing with M (cf. Figure 5D in [12]). This presents a problem when applied to the CET data. The CET data are truncated to the nearest 0.1°C, and early application of this technique encountered the problem of successive maxima in the RHS of (4). Truncation effects can be removed by adding a random uniform deviate that brackets the rounded value by half the truncation value; this will not replace lost information [13] but does allow estimation of the PDF's large-scale structure. I have smoothed the CET data by adding to each observation a uniform random deviate $\delta_i \in [-0.05, 0.05)$; this smoothing allows PDF estimation using (4) and, when truncated to the nearest 0.1°C, yields the original timeseries. Multiple random smoothings of the CET have been performed, and SE and KLD results computed from the different smoothings agree [10].

4. Central England Temperature Record

The CET is the longest observational record for surface air temperature and one of the most thoroughly studied [14]. Manley's original CET [15, 4] comprises monthly averages beginning in 1659. The daily CET [16] (Figure 1(a)) spans the period 1772–present for average temperatures (T_{avg}) and 1878–present for minimum (T_{min}) and maximum (T_{max}) temperatures. Sampling periods were 1659–2009 for monthly T_{avg}, 1772–2006 for daily T_{avg}, and 1878–2006 for (T_{min},T_{max}). All data were obtained from the British Atmospheric Data Centre [17]. CET daily and monthly temperature values are rounded to the nearest 0.1°C, with the exception of the monthly record, which has periods of low (0.5°C–1.0°C) precision for the periods 1659–1699 and 1707–1721.

The CET has a secular warming trend (Figure 1(a)) [14], which emerges after 1900. The 18th and 19th centuries are periods of relative climate stability in that there are oscillations but no overall trend in the CET [14]. The CET exhibits oscillatory behavior at multiple periods up to and beyond the century scale [18]. Figure 1(b) (1(c)) shows the daily CET average temperature PDF $\rho_{\text{full}}(T_{\text{avg}})$ ($\rho_{\text{preind}}(T_{\text{avg}})$) for the full record 1772–2006 (preindustrial era 1772–1870). These PDFs were obtained by using the technique outlined in Section 3.



(a) (b) (c)

Figure 1: CET: (a) daily T_{avg} timeseries (1772–2006), (b) full record $\rho_{\text{full}}(T_{\text{avg}})$ (1772–2006), and (c) preindustrial record $\rho_{\text{preind}}(T_{\text{avg}})$ (1772–1870)

5. CET Probability Density Functions and Informatic Variability

The time-dependent PDF $\rho(t, T_{\rm avg})$ for CET monthly averages (Figure 2(a)) was computed by using (5). The PDF is bimodal, with the lower (upper) mode steady during the 17th century at about 5°C (15°C). The upper mode is broadened and flattened around the year 1800 and reappears shortly afterward, trending gently upward through the late 20th century. The lower mode is boadened and flattened around the year 1820, reappearing shortly afterward. This mode shifts toward warmer temperatures during the 19th and 20th centuries more dramatically than does the upper mode, and the shift steepens after 1950. The early years of the sample 1659–1721 show a more detailed multimodal structure. The reason is that the optimal binning scheme chose more bins for 30-year windows during this period, a consequence of the algorithm's ability to detect truncated data

Uncertainties in $\rho(t, T_{\rm avg})$ (Figure 2(b)) were computed from (6) and show higher values throughout the temperature spectrum for the period 1659–1720, casting doubt on their associated PDF values in Figure 2(a). The ratio of the values in Figure 2(a) to those in Figure 2(b) defines a signal-to-noise (S/N) ratio (Figure 2(c)) that quantifies confidence in $\rho(t, T_{\rm avg})$. Note that little confidence (S/N < 10) is associated with any values of $\rho(t, T_{\rm avg})$ for $t \in [1659, 1721]$, except for the higher-precision period 1699–1706. The values of $\rho(t, T_{\rm avg})$ in the vicinity of its modes after the year 1721 are significant (S/N > 10).

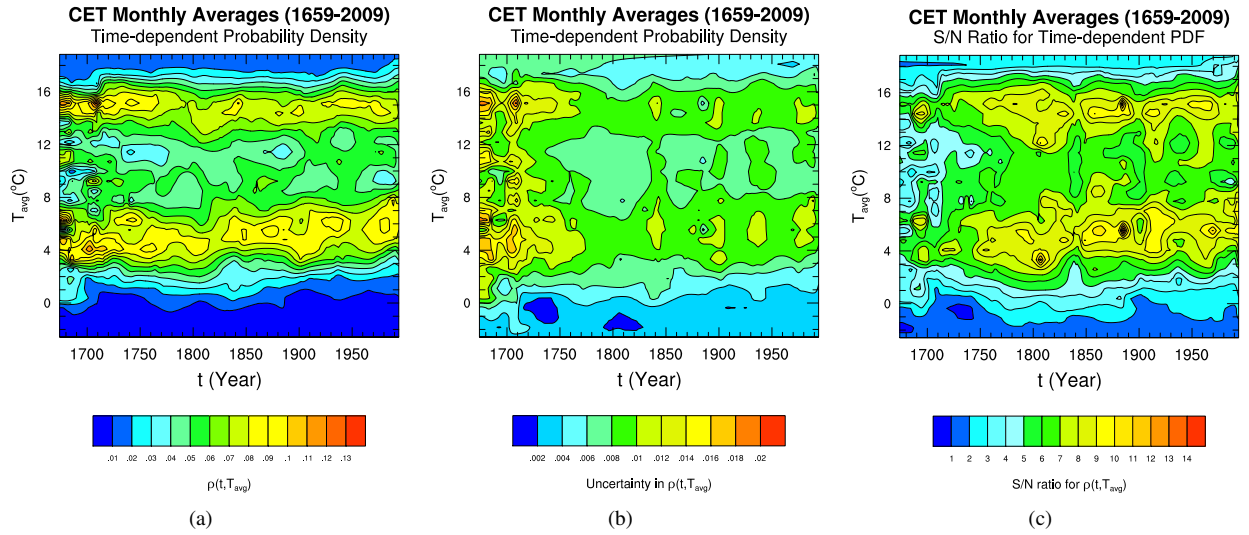


Figure 2: CET monthly $T_{\rm avg}$: (a) $\rho(t, T_{\rm avg})$, (b) uncertainties in $\rho(t, T_{\rm avg})$, and (c) signal-to-noise ratio for $\rho(t, T_{\rm avg})$.

PDFs for the daily CET are presented in Figure 3. The PDF $\rho(t, T_{\rm avg})$ (Figure 3(a)) is bimodal, with the upper (lower) mode centered near 7°C (14°C). The lower mode is less pronounced before 1830. The warming period 1910–1950 is present with a narrowing and upward shift of the lower mode, which stabilizes for the period 1950–1975 before again shifting upward and broadening. This alternating strengthening (weakening) of the upper mode while the lower mode is weak (strong) is a pattern that is seen during the period 1790–1950 and has a wavelike character. If this pattern is considered a wave, its period τ may be estimated by measuring the time interval between peaks in the upper mode, or the width of the "island" in the lower mode; both approaches yield a period of $\tau \approx 125$ years. The S/N ratios for $\rho(t, T_{\rm avg})$ (Figure 3(d)) show significant values broadly in the band $-1°{\rm C} < T_{\rm avg} < 21°{\rm C}$, with the central area around the modes highly significant (S/N> 20). Within this region, PDF isopleths shift upward dramatically after 1980. The PDF $\rho(t, T_{\rm min})$ (Figure 3(b)) shows a single mode. This mode is broad around 1930, with width $\Delta T_{\rm min} \approx 7°{\rm C}$ and center $T_{\rm min} \approx 5.5°{\rm C}$. After 1930, the mode narrows dramatically to $\Delta T_{\rm min} \approx 4°{\rm C}$, with a slight warming up until 1950.The mode stabilizes and broadens near 1960, with width $\Delta T_{\rm min} \approx 7.5°{\rm C}$, and its center shifts upward to $T_{\rm min} \approx 6.0°{\rm C}$. During the period 1960–1980, the mode broadens further to $\Delta T_{\rm min} \approx 8.5°{\rm C}$, and its center remains stationary. After 1980, the mode narrows, with its lower boundary PDF isopleth moving upward much more dramatically—a shift of nearly 2.0°C—than its upper counterpart. The mode's center shifts upward to 8°C. This narrowing and rebroadening of the mode may be considered a wave structure in $\rho(t, T_{\rm min})$; if viewed that way, it has

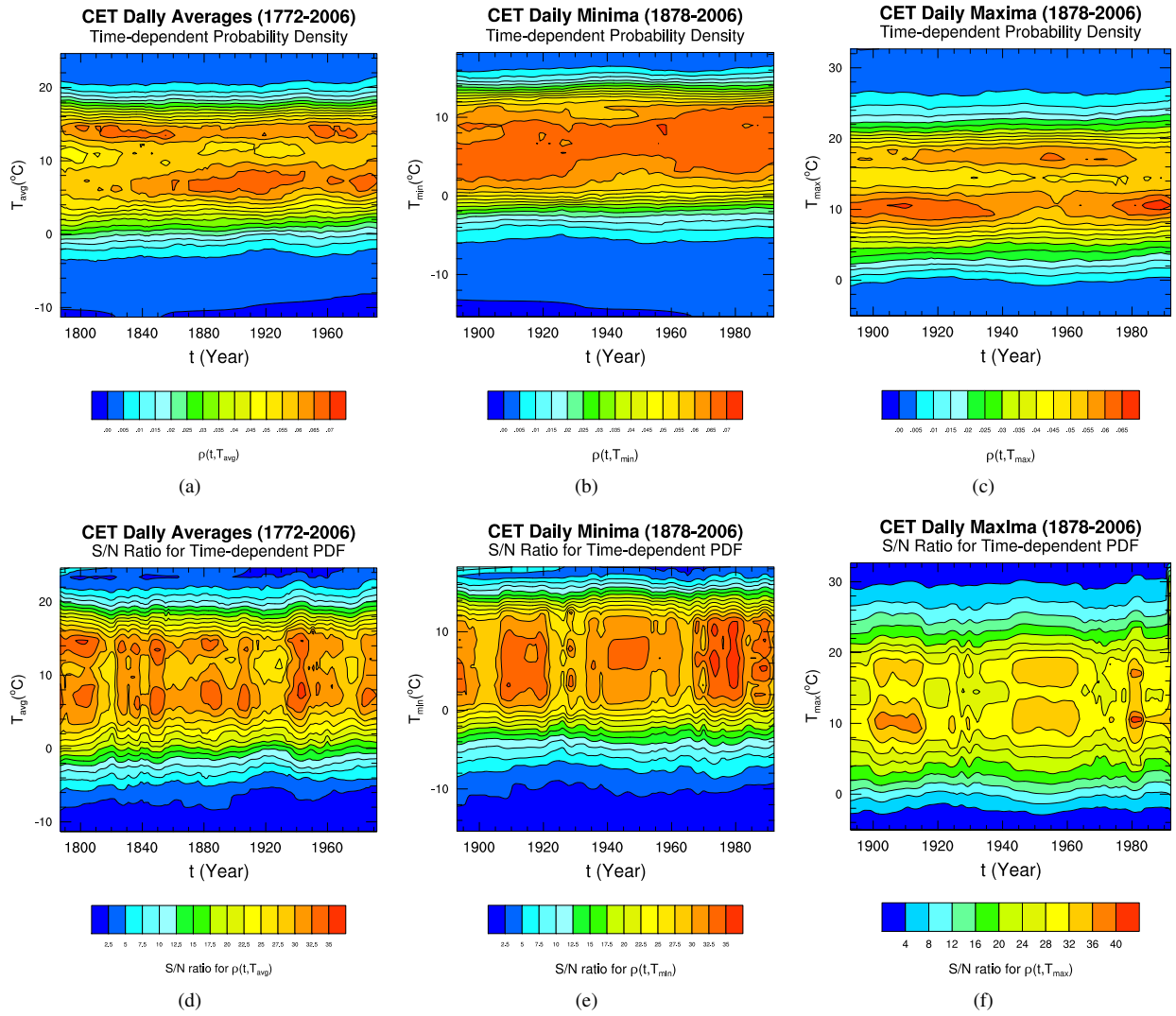
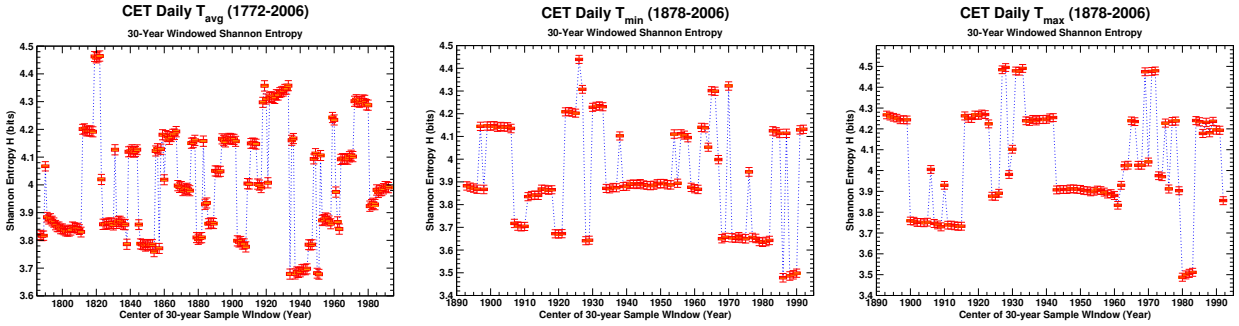


Figure 3: Daily CET PDFs: (a) $\rho(t, T_{\text{avg}})$ 1772–2006, (b) $\rho(t, T_{\text{min}})$ 1878–2006, and (c) $\rho(t, T_{\text{max}})$ 1878-2006. Signal-to-noise ratio in time-dependent PDFs for (d) T_{avg}, (e) T_{min}, and (f) T_{max}.

a period $\tau \approx 70$ years, based on measurement between the centers of the broad portions of the mode. S/N values in Figure 3(e) for $\rho(t, T_{\text{min}})$ show significance for $T_{\text{min}} \in [-5, 15]$, high significance for $T_{\text{min}} \in [-2, 13]$, abd highest significance values clustered about $T_{\text{min}} \in [0, 11]$, covering the PDF's shifting mode. Within this high S/N zone, PDF isopleths in Figure 3(b) show weak variation superimposed on a slow warming trend of approximatly 1.0°C for the period 1878–2006. The PDF $\rho(t, T_{\text{max}})$ (Figure 3(c)) is bimodal, though neither mode is present for the full record. The upper mode has width and center ($\Delta T^U{}_{\text{max}} \approx 3$°C, $T^U{}_{\text{max}} \approx 17.5$°C) and is most evident during the period 1910–1980. Its weakness before 1910 appears to be due to broadening of the midrange of T_{max}. The disappearence of this mode after 1980 is caused by midrange broadening and fattening of the high-T_{max} tail. The lower mode has width and center ($\Delta T^L{}_{\text{max}} \approx 4$°C, $T^L{}_{\text{max}} \approx 10.5$°C). It is most pronounced during the periods 1893–1937 and 1973–2006. It disappears briefly around 1955, just when the upper mode is at its strongest. Taken together, these modes suggest an oscillation in $\rho(t, T_{\text{max}})$ with a period of $\tau \approx 80$ or $\tau \approx 70$ years if estimated from the peaks in the lower mode or width of the upper mode, respectively. All these results are significant for 1°C $< T_{\text{max}} <$ 24°C and highly significant for 4°C $< T_{\text{max}} <$ 21°C (Figure 3(f)).

The oscillatory nature of the PDFs is striking and may be related to previously known oscillations found in the

monthly CET by Benner [18]. Benner estimated power spectra for the CET using four techniques: the fast Fourier transformation (FFT), the Lomb-Scargle periodogram (LSP), singular spectrum analysis (SSA), and the global wavelet spectrum (GWS). He found many different oscillation periods in the CET's spectrum, and three of these techniques—FFT, LSP, and GWS—identified periods near those in the CET daily $\rho(t, T_{\rm avg})$, $\rho(t, T_{\rm min})$, and $\rho(t, T_{\rm max})$. The 125-year oscillation that appears in $\rho(t, T_{\rm avg})$ may be related to the long-period oscillation in the monthly CET $T_{\rm avg}$ timeseries that has $\tau_{\rm FFT} = 113$, $\tau_{\rm LSP} = 112.97$, and $\tau_{\rm GWS} = 108.53$ years. The 70–80-year oscillation that appears in $\rho(t, T_{\rm min})$ and $\rho(t, T_{\rm max})$ may be related to the interdecadal oscillation in the monthly CET $T_{\rm avg}$ timeseries that has $\tau_{\rm FFT} = 67.8$, $\tau_{\rm LSP} = 67.78$, and $\tau_{\rm GWS} = 69.77$ years, possibly the Atlantic Multidecadal Oscillation (AMO) [19].



(a) (b) (c)

Figure 4: Thirty-year windowed Shannon entropy H for CET daily (a) $T_{\rm avg}$ (1772–2006), (b) $T_{\rm min}$ (1878–2006), and (c) $T_{\rm max}$ (1878–2006).

SE timeseries $H(t)$ were computed from the $\rho(t, T)$ by using (2). Uncertainty quantification was performed by computing H using Monte Carlo–generated ensembles of 10,000 "neighboring" PDFs defined by (5) and (6) (Figure 4). The box-whisker symbols are defined with whiskers corresponding to the 1st and 99th percentiles, box edges to the 25th and 75th percentiles, and box center to the median. The dramatic up/down jumps in H with respect to time are thus unlikely to be numerical artifacts. The timeseries of $H(T_{\rm avg})$ (Figure 4(a)) shows a pronounced peak at its maximum value shortly before 1820, when the lower mode was weak to nonexistent; this broadening would dramatically lower local values of the PDF and thus create larger logarithmic terms in (2). The periods with low (high) values of $H(T_{\rm avg})$ frequently correspond to times when both the upper and lower modes in $\rho(t, T_{\rm avg})$ are present (absent or weak), but this correspondence is not complete. The SE $H(T_{\rm max})$ (Figure 4(c)) are equally hard to interpret. Each jump or dip in $H(T_{\rm max})$ indicates some structural change in $\rho(t, T_{\rm max})$, but identifying the responsible feature(s) is difficult. The simplest of the SE plots to interpret is that of $H(T_{\rm min})$ (Figure 4(b)) because $\rho(t, T_{\rm min})$ is practically unimodal. The broad high value during 1895–1905 appears to be due to splitting of the mode. Narrow local maxima $H(T_{\rm min})$ are present at other times when the mode is split, namely, shortly before 1920 and 1930 and in the early 1980s. A wide, almost uninterrupted trough in $H(T_{\rm min})$ for 1930–1950 appears to be caused by the sharpening of the mode. By contrast, an even lower trough during 1965–1982 coexists with the mode being at its widest. Although the SE is clearly a sensitive integrated measure, its integrative nature obscures which features change its value. Thus, **Q2** is answered, but the answer's utility is unclear.

Figures 5 (a,c,e) collectively answer **Q3** for the CET monthly $T_{\rm avg}$ and daily ($T_{\rm avg}$,$T_{\rm min}$, $T_{\rm max}$). Figure 5(a) shows the KL gain from any 30-year sample window's PDF, to predict PDFs derived from the full sample (1659–2009; solid line) and the preindustrial era (1659–1869; dotted line). From previous discussion, the high KLD values for 1659–1750 are suspect because of truncation effects during the period 1659–1721. KLD values for the period 1750–1890 are relatively low, with some oscillatory behavior. After 1890 there are dramatic jumps in the KLD to the whole-record PDF and even more dramatic increases in the KLD to the preindustrial PDF. Some drop occurs during the stabilization period 1950–1975, followed by even more dramatic increases after 1975. Late in the 20th century the KL gain to the preindustrial PDF is higher than for any other part of the CET record, signifying strong structural change in the PDF since the preindustrial era. The KL gains from the late 20th century PDFs $\rho(t, T_{\rm avg})$ are also high. Overall, the KL gain identifies much of the 20th century's time-evolving climate PDFs as distinctly weaker in their ability to predict the long-scale record, and thus structurally fundamentally different. The daily $T_{\rm avg}$ (Figure 5(c)) shows a similar degradation in skill through increasing KLD values during the 20th century. The representativeness

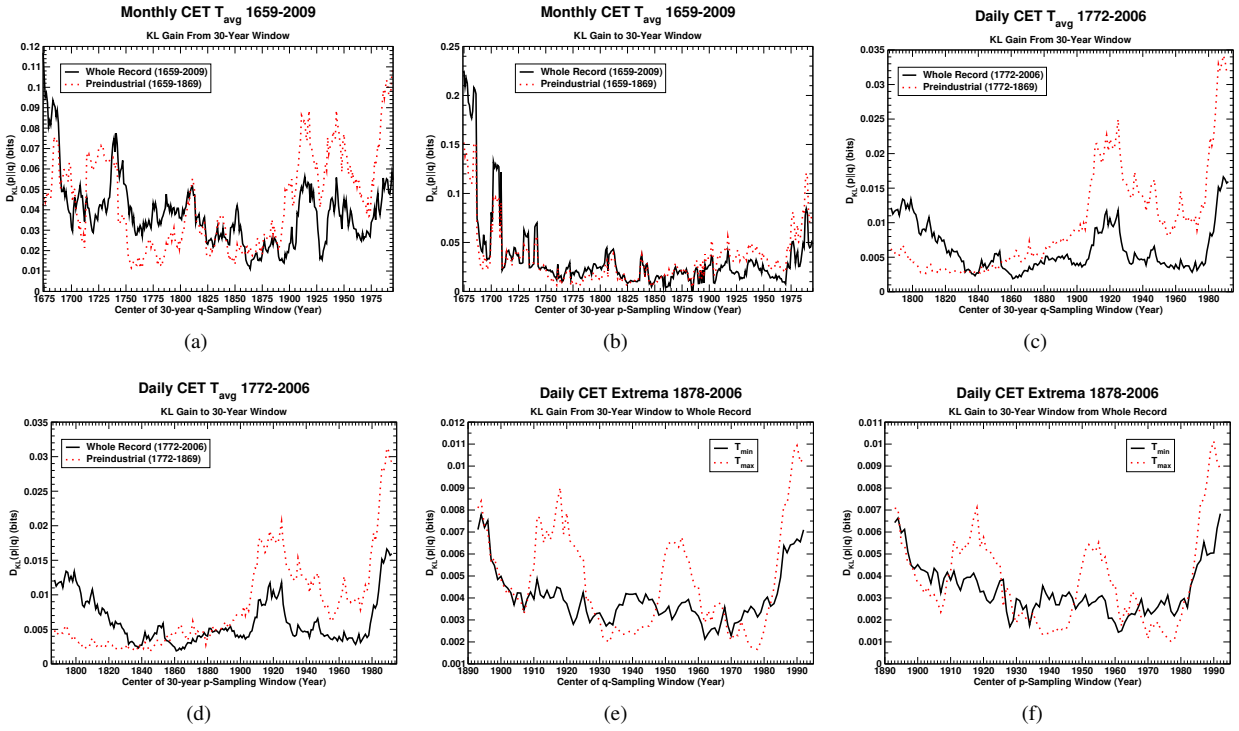
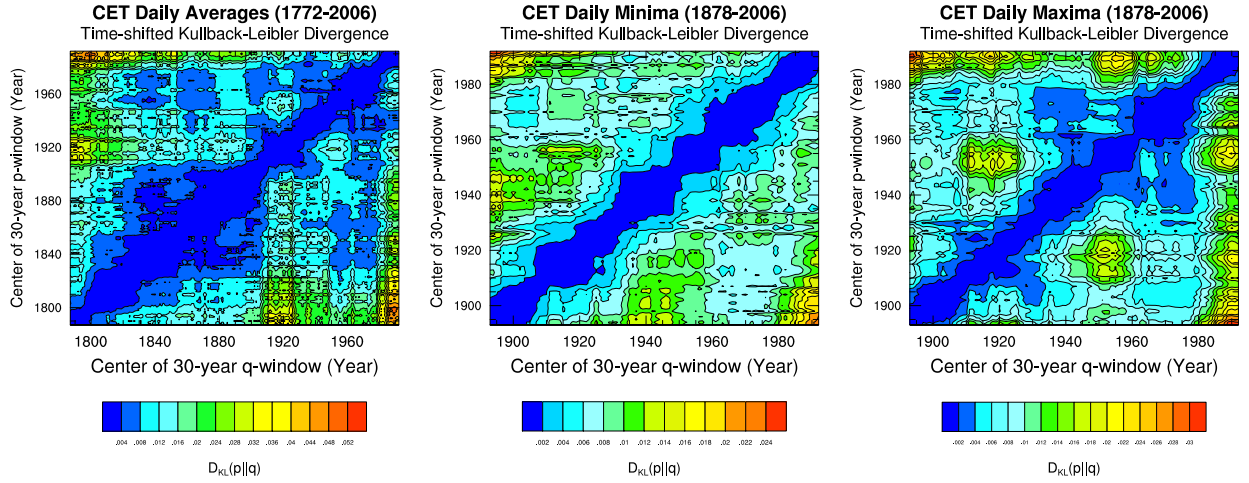


Figure 5: CET Kullback-Leibler divergences. Divergences from 30-year sampling windows to larger records for (a) monthly CET T_{avg}, (b) daily CET T_{avg}, and (c) daily CET T_{min} and T_{max}. Divergences from larger record to 30-year sampling windows for (d) CET monthly T_{avg}, (e) CET Daily T_{avg}, and (f) CET daily T_{min} and T_{max}.

of the CET extreme values of their whole record 1878–2006 (Figure 5(e)) show marked contrast between T_{min} and T_{max}. The KL gain from $\rho(t, T_{\text{max}})$ is highly cyclic, with strong peaks at 1895, 1918, 1955, and 1990, which suggest a connection to the oscillation present in $\rho(t, T_{\text{max}})$. The 1990 peak is the highest, confirming that $\rho(1990, T_{\text{max}})$ is the most distantly related PDF to the complete record's PDF, again suggesting that recent T_{max} climatology is distinctly different from the century that preceded it. The KL gain from $\rho(t, T_{\text{min}})$ to the complete record's $\rho(T_{\text{min}})$ has high peaks around 1895 and 1995, with weaker variability imposed on a trough spanning 1905–1980. Again, this suggests any 30-year period centered on one of the years 1905–1980 is much more representative than the peaks at either end of the record. For the period 1980–1995 the KL gain required to predict $\rho(T_{\text{min}})$ from $\rho(t, T_{\text{min}})$ is nearly as high as the highest value seen in 1895 and suggests that recent T_{min} climatology is relatively alien compared with the full record.

Figures 5 (b,d,f) collectively answer **Q4** for the CET monthly T_{avg} and daily (T_{avg} ,T_{min}, T_{max}), that is, the amount of additional information needed to predict $\rho(t, T)$, given the full record's PDF $\rho(T)$. The monthly T_{avg} show KL gain to any 30-year sample-generated $\rho(t, T_{\text{avg}})$ from the full and preindustrial PDFs to be large for the early, deprecated part of the record and the period after 1975. A long trough covers 1740–1975, with some weaker variability superimposed on it (Figure 5(b)). This is true for the full-sample and preindustrial PDFs. Significantly, the strongest significant values for these KL gains lie in the period after 1975, again demonstrating that recent climatology is odder than any other 30-year period. For the daily T_{avg} the KL gain to $\rho(t, T_{\text{avg}})$ from the full-record and preindustrial PDFs (Figure 5(d)) identifies the period after 1980 as being the most distantly related to the full and preindustrial records. The gains necessary to predict recent sample PDFs from the full and preindustrial records are dramatically greater than those seen for the 1910–1950 warming period. It is no surprise that the information gain required to predict $\rho(t, T_{\text{avg}})$ from the preindustrial sample rises steadily from 1870 forward, a consequence of the p- and q- sampling windows becoming disjoint after 1900 and then being separated by increasing time lags up to the present. The respite from this rising trend is associated with the 1950–1975 cooling/stabilization period. The information gain from the PDF of the full daily CET extrema record 1878–2006 to any 30-year windowed PDF (Figure 5(f)) is structurally similar to Figure 5(e) but with slightly lower peaks. The reason is that the q-window used to generate the full PDF contains the

reference PDF's p-window as a subset. Note again that the full record's PDF has the most trouble predicting PDFs generated from windows centered on years after 1985, identifying the climatology of daily CET extreme temperatures as distinctly odd with respect to the full observational record.



(a) (b) (c)

Figure 6: Time-shifted Kullback-Leibler divergences for the daily CET: (a) T_{avg} (1772–2006), (b) T_{min} (1878–2006), and (c) T_{max} (1878–2006).

The answer to **Q5** can be found by computing time-shifted KLD values that compare each 30-year window to every other 30-year window present in the observational record. Analysis of these results will provide clues to whether we can answer **Q6** as well. Time-shifted KLD values for the daily CET are shown in Figure 6. These plots have as ordinate (abcissa) the center year t_q (t_p) of a 30-year sampling window used to generate the PDF $\rho(t_q, T)$ ($\rho(t_p, T)$). The value plotted at the point (t_q, t_p) is $D_{KL}(p \parallel q)$; this value is zero on the line $t_q = t_p$ because $D_{KL}(p \parallel p) = 0$. For a fixed value of t_q, points vertically above (below) the point (t_q, t_q) signify the ability of this windowed PDF to predict future (past) climate PDFs. The CET daily T_{avg} time-shifted KLD results (Figure 6(a)) reveal a block-diagonal structure comprising multiple low-KLD-value regions; these correspond to periods of relative climate PDF stability in that any 30-year window from this range can, with relatively high accuracy, predict another 30-year windowed PDF in this range. Stable time ranges identified are 1800–1850, 1830–1910, 1910–1950, and 1950–1975. Note that some of these intervals touch or overlap; overlapping intervals signify a gentler transition in terms of PDF structure, while intervals that merely touch indicate a more abrupt change in PDF structure. These results are consistent with Figure 3(a). Off-diagonal blocks of low KLD values indicate periodicity expressed as similarities from windowed PDFs from one period versus another. In particular, the intervals 1950–1970 and 1830–1870 generate similar PDFs. Note further that only a modest diagonal band is found after 1980, indicating relatively rapid change in $\rho(t, T_{\text{avg}})$ during this period. Also note that the highest time-shifted KLD values are in vertical (horizontal) bands defined by $t_q > 1980$ ($t_p > 1980$), indicating that climate PDFs centered after 1980 are distinctly different from pre-20th century climate. Time-shifted KLD values for the CET daily T_{min} record (Figure 6(b)) show a narrow block-diagonal structure, indicating a series of overlapping short periods of relative PDF stability. Strong contrasts between the warming period of the early 20th century and mid-century cooling/stabilization period that follow it are evident in relatively strong KLD values. The sampling period after 1980 again appears distinctly different from most other time periods. The time-shifted KLD values for T_{max} (Figure 6(c)) show a diagonal band whose width varies from 5 to 15 years, with some block-diagonal structures associated with the periods 1910–1930, 1940–1960, and 1960–1975. These periods are times of relatively slow change in $\rho(t, T_{\text{max}})$ Off-diagonal low-KLD-value blocks indicate periodicity with the intervals 1930–1950 and 1960–1980 appearing related. Off-diagonal high-KLD blocks indicate a strong divergence between the statistics associated with the sampling periods centered on the intervals 1910–1925 and 1945–1960. Note the strong divergence between PDFs associated with sampling windows centered on years after 1980 and the rest of the record, again signaling a fundamental shift in the structure of $\rho(t, T_{\text{max}})$.

6. Conclusions and Future Work

An exploratory data analysis/information theoretic framework to analyze climate variability has been developed and applied to the Central England Temperature record. Viewing the 30-year windowed PDFs has provided, at a glance, deep insight into the evolution of CET monthly and daily average and daily extreme temperatures. The time-dependent PDFs are consistent with and add new detail to the CET's known warming trend. These PDFs also exhibit oscillatory behavior that may be connected to known interdecadal and century-scale oscillations present in the CET monthly average timeseries. The KLD-based measures of representativeness and oddness put the climate of the past 30 years in stark contrast with respect to the full and preindustrial observational records. The climatology of CET average, maximum, and minimum temperatures is distinctly different from that of previous observed times. Time-shifted KLD metrics have identified previously known periods of relative climatic stability. The metrics cast the climate of recent decades as distinctly different and changing rapidly with respect to the past century's climate.

The results reported here are preliminary. Near-term areas of future investigation include using other density estimation techniques to verify these results, applying spectral techniques to the time-dependent PDFs to search more thoroughly for periodicity, developing automatic feature detection schemes to identify periods of climatic stability and change, and applying these techniques to larger observational and model-generated data sets. The long-term goal of this work is to determine whether equations of evolution for time-dependent climate PDFs—something akin to (1)—may be reliably estimated from timeseries data and whether such empirical models have any predictive power.

Acknowledgment

This work was supported by the U.S. Department of Energy, under Contract DE-AC02-06CH11357.

References

[1] M. Hulme, S. Dessai, I. Lorenzoni, D. R. Nelson, Unstable climates: Exploring the statistical and social constructions of 'Normal' climate, Geoforum 40 (2009) 197–206. doi:10.1016/j.geoforum.2008.09.010.
[2] International Meteorological Organization, Proceedings of the Meetings in Danzig and Warsaw, 29–31 August and 12 September 1935, Secretariat of the IMO, Leyden, 1937.
[3] K. Hasselmann, Stochastic climate models, Part I: Theory, Tellus 28 (6) (1976) 473–485. doi:10.1111/j.2153-3490.1976.tb00696.x. URL http://dx.doi.org/10.1111/j.2153-3490.1976.tb00696.x
[4] G. Manley, Central England temperatures: Monthly means 1659 to 1973, Quarterly Journal of the Royal Meteorological Society 100 (425) (1974) 389–405.
[5] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423.
[6] T. M. Cover, J. A. Thomas, Elements of Information Theory, 2nd Edition, Wiley-Interscience, New York, 2006.
[7] T. DelSole, M. Tippett, Predictability: Recent insights from information theory, Reviews of Geophysics 45 (2007) RG4002.
[8] J. Shukla, T. DelSole, M. Fennessy, J. Kinter, D. Paolino, Climate model fidelity and projections of climate change, Geophysical Research Letters 33 (2006) L07702.
[9] J. W. Larson, Information-theoretic strategies for quantifying variability and model-reality comparison in the climate system, in: R. S. Anderssen, R. D. Braddock, L. T. H. Newham (Eds.), Proceedings of the 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, 2009, pp. 2639–2646.
[10] J. W. Larson, Can we define climate using information theory?, IOP Conference Series: Earth and Environmental Science 11 (1) (2010) 012028.
[11] D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, New York, 1992.
[12] K. H. Knuth, Optimal data-based binning for histograms, http://arxiv.org/abs/physics/0605197 (2006).
[13] K. H. Knuth, J. P. Castle, K. R. Wheeler, Identifying excessively rounded or truncated data, in: A. Rizzi, V. Maurizio (Eds.), Proceedings of the 17th meeting of the International Association for Statistical Computing—European Regional Section: Computational Statistics (COMPSTAT 2006), Springer, 2006.
[14] P. D. Jones, M. Hulme, The changing temperature of central England, in: M. Hulme, E. Barrow (Eds.), Climates of the British Isles, Present, Past, and Future, Routledge, London, 1997, pp. 173–196.
[15] G. Manley, The mean temperature of central England 1698–1952, Quarterly Journal of the Royal Meteorological Society 79 (340) (1952) 242–261.
[16] D. E. Parker, T. P. Legg, C. K. Folland, A new daily central England temperature series 1772–1991, International Journal of Climatology 12 (1992) 317–342.
[17] British Atmospheric Data Centre, Hadley Centre, UK Meteorological Office, Historical Central England Temperature (CET) Data, http://badc.nerc.ac.uk/data/cet/ (2009).
[18] T. C. Benner, Central England temperatures: long-term variability and teleconnections, International Journal of Climatology 19 (1999) 391–403.
[19] M. E. Schlesinger, N. Ramankutty, An oscillation in the global climate system of period 65-70 years, Nature 367 (6). doi:10.1038/367723a0.

Government License